Special issue

Collection and sharing of genomic and health data for research purposes: Going beyond data collection in traditional research settings

Mahsa Shabani *

ABSTRACT: In the recent years collection of health and genomic data for biomedical research purposes has been expanded beyond traditional research settings. In doing so, various online tools and platforms are being utilized to collect data from various sources including Electronic Health Records, mHealth applications, disease registries and patient generated databases. While there is relatively higher certainty regarding the legal grounds for processing health and genomic data in the traditional research setting, the questions remain about the applicable legal framework when collecting data from other sources. In addition, given the diverse nature of collected data, adhering to traditional care-research distinction to determine the applicable legal requirements is confronted with complexities. This is particularly the case when data collected in the care setting are being later used for research purposes. In this article, we discuss the challenges associated with governance of processing data collected outside research settings and underline the steps should be taken to ensure conformity of such data processing by the applicable data protection regulations.

KEYWORDS: Genomic data; biomedical research; mHealth; Real-World Data; GDPR

SUMMARY: 1. Introduction – 2. Data collection via mHealth applications and online platforms – 3. Data protection in the context of mHealth applications and online platforms – 4. Collection of Real-World Data (RWD) – 5. Data Protection framework for processing RWD – 5.1 Source of the initial data collection and a question of further processing of data – 5.2. Legal requirements for using RWD for research purposes – 6. Path forward.

1. Introduction

Genomic research requires access to a large scale of research and clinical data in order to improve the statistical power of the databases and assist finding similar cases. In the recent years, a need for access to large scale of data has led to increasing support for data sharing among researchers across the world. In particular, researchers who are funded by public funding are strongly recommended or in some instances mandated to share their genomic data through public genomic databases such as dbGaP (the database of Genotypes and Phenotypes) and

* *Metamedica, Faculty of Law and Criminology, Ghent University, Ghent, Belgium. Mail: mahsa.shabani@ugent.be.* *This research has been partially supported by the research funding from FWO Flanders-Quebec project. The article was peer-reviewed by the editorial committee.*

EGA (European Genome-phenome Archive)[1]. This way, the data generated through public money can be put in optimal use and other researchers across the world can have the opportunity to run downstream analysis on the existing databases.

In addition, collection of health and genomic data outside the traditional research setting has received increasing attention. One of the important sources of such data is related to patients and individual-generated data which can be collected via various mHealth related applications and online platforms. This way, individuals could contribute health and genomic data either by uploading their genomic test results or submitting other health related information through filling questionnaires or self-measurement of various variables. This would also facilitate collection of so-called real-world data (RWD), namely data which are routinely collected outside a controlled research environment. As it has been stated in a recent report by the Duke Margolis Center for Health Policy: "Mobile health (mHealth) apps and wearables, particularly those that collect patient- and consumer-generated health data, can fill some of […] data gaps by providing real-world, more meaningful, high frequency, and/or longitudinal data." [2]

Real World Data have attracted an increasing attention in recent years due to the fact that evidence provided by traditional clinical research often fails to answer patients', physicians' and healthcare decision-makers' questions about real-world practices and outcomes. This is specially of interest in the context of clinical trials, as it has been expressed in a statement issued by Roche on Access to and Use of RWD: "While clinical trials focus on ensuring that valid causal conclusions can be drawn between intervention and effect, RWD are seen as a potentially rich and underutilized source to generate insight as to how approved diagnostics systems and medicines affect outcomes for patients under real world conditions."[3]

The collection and processing of health data for research purposes outside traditional research setting including using consumer health applications and other mobile devices, raises a number of privacy and data protection issues. Notably, privacy concerns are being intensified in the context of genetic data, as full anonymization of genetic data is not possible, owing to the nature of genetic data which contains unique identifiers about the individuals. In Europe, collection and use of personal data must be compliant with the EU General Data Protection Regulation (GDPR), and other applicable national regulations.

In the framework of the GDPR, which has been a main data protection legal instrument in the EU since its implementation in 2018, personal data is defined as any information relating to an identified or identifiable natural person. Therefore, processing data that have been irreversibly de-identified may fall outside the scope of the relevant personal data protection regulations. The GDPR does not specify the exact identifiers which need to be removed in order to render data non-identifiable, but notes that "to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another

[1] M. Shabani M, B. Knoppers, P. Borry. *From the principles of genomic data sharing to the practices of data access committees*, in *EMBO Molecular Medicine*, 7, 2015, 507-509.

[2] Duke Margolic Center for Public Policy. Mobilizing mHealth Innovation for Real World Evidence Generation. Available online at: https://healthpolicy.duke.edu/sites/default/files/2020-03/duke-margolis_mhealth_action_plan.pdf.

[3] Roche, *Roche Position on Access to and Use of RWD*, 2019, available online at: https://bit.ly/32p5JWE.

ISSN 2284-4503

person to identify the natural person directly or indirectly."[4] Other regulations concerning health data, such as HIPAA (Health Insurance Portability and Accountability Act) in the USA, take a different approach, namely by listing specific identifiers that should be removed from datasets in order to de-identify them.[5]

The GDPR foresees, among other things, specific provisions regarding processing sensitive personal data, including health and genetic data, and provisions regarding processing data for scientific research purposes. In particular, it should be ensured that data processing is based on one of the recognized legal bases under the applicable data protection regulations. In addition, it should be ensured that adequate organizational and technical measures are adopted to mitigate the risks of re-identifiability of data and privacy breaches when processing individuals' data.[6]

In this manuscript, first, we provide an overview of emerging data collection approaches for biomedical research purposes outside traditional research setting, including collection of data via mHealth applications and online platforms, and collection of real-world data. In the following, we will discuss the associated legal concerns related to collection and processing of such data and analyze how the existing data protection framework would apply to the data collection and use outside traditional research setting, and what would be the future approaches regarding applicable legal framework for such data collection and data reuse.

2. Data collection via mHealth applications and online platforms

Smartphone applications for health are being increasingly used as a platform for collecting mass volumes of crowdsourced personal health data. Smartphones, for example, record and process numerous health measurements, including behavioral measures, clinical data, and health related symptoms. More than five billion people around the world own some kind of mobile device and in advanced economies, rates of smartphone ownership average nearly 80%.[7] A majority of smartphone owners have used their devices to access, record, or track data relevant to their health.[8] The popularization of personal health data monitoring, the increasing ubiquity of smartphone applications with health-related functions, and the rapidly improving technological capacities of mobile devices, have led to unprecedented opportunities for harnessing crowdsourced health and genomic data to expand biomedical knowledge. At the same time, the increasing popularity of consumer genetics products is permitting individual result processing while also facilitating the

[4] Recital 26, GDPR.

[5] M. SHABANI, L. MARELLI, *Re-identifiability of genomic data and the GDPR Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation*, in *EMBO reports*, 20, 2019, e48316.

[6] Article 89 (2), GDPR.

[7] L. SILVER, *Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally*, *Pew Research Center*, 2019, available online at: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/.

[8] A. EDWARDS, *mHealth: Healthcare Mobile App Trends in 2019,* in *Ortho Live,* 2019, available online at: https://www.ortholive.com/blog/mhealth-healthcare-mobile-app-trends-in-2019.

sharing of such information with clinicians and health researchers,[9] potentially complementing and enhancing the utility of data collected on mobile device platforms.

The collection of genomic and health data via mHealth applications has been also considered as a citizen science activity, whereby the general public is directly involved in collection of data and conducting scientific research.[10] In the context of genomic research, this has been enabled by availability of Direct-to-Consumer genetic testing services which allow individuals to have access to their raw data and share that for research purpose. To name a few examples of such projects, Citsci.org, Open Humans.org, and the platform created by the for-profit company PatientsLikeMe are currently providing platform for citizen science-based data collection.[11] This will promote easier participation of the individuals in the research and facilitate efficient access to individual-generated data.

A study conducted by Talwar and colleagues have systematically reviewed the existing genetics and genomics apps in the market. They concluded that "while the majority of the apps served as references or resources (i.e., providing general genetics/genomics information and/or tutorials), some apps provided lifestyle recommendations to the general public, mostly based on the DTC genetic test results. Specifically, using integrated schemes and algorithms, those apps could provide an interpretation of genetic test results and then offer personalized recommendations regarding nutrition and physical activity." [12] This study identified eighty-eight genetics and genomics related apps, although it is not clear how many of these apps also collect and reuse users' data for research purposes. As developing and improving apps require access to a large scale of data, it is expected that the collected data from the users to be considered as a valuable resource for the app developers. This feature is a prominent aspect in the context of AI-driven technologies which use machine learning methods for finding a pattern in data.

In addition, other initiatives have been emerging which aim to collect personal genomic data from individuals in exchange for various financial incentives. The examples are Luna DNA, Nebula Genomics and ENCRYPGEN which invite individuals to upload their DNA data to be used for various research and clinical purposes by the interested parties. To make such data sharing by the individuals fair, they offer various incentives in exchange for data, including returning a free DNA report, DNA tokens or shares. [13] Although these platforms allegedly are utilizing innovative ways to collect personal genomic data, concerns remain regarding privacy and ownership, and compatibility of these approaches with the research ethics principles. We will discuss some of these concerns in the following part.

---

[9] M CABELL *Jonas* et al*. Physician Experience with Direct-To-Consumer Genetic Testing in Kaiser Permanente*, in *Journal of Personalized Medicine,* 9:4, 47, 2019.

[10] E. HAFEN. *Personal Data Cooperatives – A New Data Governance Framework for Data Donations and Precision Health*, in: J. KRUTZINNA, L. FLORIDI (eds) *The Ethics of Medical Data Donation. Philosophical Studies Series*, in *Springer*, 137, Cham, 2019, https://doi.org/10.1007/978-3-030-04363-6_9

[11] M. MAJUMDER, A. MCGUIRE. *Data Sharing in the context of Health-related Citizen Science*, in *Journal of Law, Medicine, and Ethics*, 2020, https://doi.org/10.1177%2F1073110520917044.

[12] D. TALWAR. *Characteristics and quality of genetics and genomics mobile apps: a systematic review*, in *European Journal of Human Genetics,* 27 (6), 2019, 833-840.

[13] E. AHMED, M. SHABANI. *DNA Data Marketplace: An Analysis of the Ethical Concerns Regarding the Particopation of the Individuals,* in *Frontiers in Genetics, 2019*, https://doi.org/10.3389/fgene.2019.01107.

3. Data protection in the context of mHealth applications and online platforms

Protecting the rights of the individuals on their personal data including their right to privacy is a cornerstone in collection, storage, uses and sharing of health-related and genomic data. Collection of data from the individuals for research purposes may raise special data protection and privacy concerns. First, when collecting individual-generated data via various mHealth applications or other health data from individuals via online platforms, it is essential to ensure that individuals are fully aware of initial and secondary data use purposes. As we have mentioned in the previous section, it is likely that the data collected from the individuals via genomics and genetics related apps to be used for further development of the apps or improvement of the existing versions, or other relevant research purposes. According to the GDPR, the data subjects have a right to receive transparent information regarding how their data are being processed and for which purposes.[14] Often mhealth applications and online platforms develop privacy policies in which they include information regarding purposes of data processing and potential secondary uses, such as for research purposes. The previous investigations have shown that often individuals do not read long privacy policy documents or are not always fully informed about the subsequent changes to the privacy policy.[15] As a result, the adequacy of the privacy policies in ensuring transparency of the data processing has been questioned at times.

Second, the proposed models of DNA data marketplaces raise an array of concerns related to adequate legal and policy framework for protection of privacy and ownership rights of the individuals when exchanging their data for free test reports, DNA tokens, shares, and the like. Notably, offering compensation for donating data for research purposes has been traditionally considered questionable under research ethics principles, due to the concerns about undue influence on individuals and potentially rendering their consent invalid.[16] In addition, in view of a legal vacuum for data ownership in many jurisdictions, it remains to be seen how monetary value of data will be evaluated in a fair and legally valid manner and to be legally protected in case of future disputes. In terms of privacy, some of these emerging platforms are suggesting adopting new technologies such as blockchain to increase the security of the data processing in a decentralized manner. Blockchain is an emerging technology of a decentralized, digitized database medium and a public ledger of all transactions in the network.[17] Blockchain-based solutions have recently gained popularity in the context of genomic and health data sharing, with the promise of improving data access, patient empowerment, and improved interoperability. [18] That said, the implementation of blockchain technology in the context of genomic and health data sharing is still in its infancy and it remains to be seen how far this can address the data protection and data access governance concerns.

[14] Art 12 & 13, GDPR.

[15] N. STEINFELD. "*I agree to the terms and conditions": How do users read privacy policies online? An eye-tracking experiment. Computers in Human Behavior*, 55 (Part B), 2016, 992-1000.

[16] E. AHMED, M. SHABANI. D*NA Data Marketplace: An analysis of the ethical concerns regarding the participation of the individuals",* in *Frontiers in Genetics,* 2019, https://doi.org/10.3389/fgene.2019.01107

[17] H. OZERCAN, I., ILERI, E. AYDAY, C. ALKAN, *Realizing the potential of blockchain technologies in genomics*, in *Genome Res*, 28, 2018, 1255–1263, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6120626/ .

[18] M. SHABANI. *Block-chain based platforms for genomic data sharing: a decentralised approach in response to the governance problems?,* in *JAMIA*, 2018, https://doi.org/10.1093/jamia/ocy149.

4. Collection of Real-World Data

Another type of health data which is of interest for biomedical research is so called real-world data (RWD) which are often collected outside controlled research setting. The rise of interest in RWD is driven by the increasing need for evidence in specific populations, such as comorbid or multi-treated people. In addition, RWD may represent the only source of information in some fields of special interest, e.g., rare diseases. RWD can also allow investigation of unanticipated, uncommon or long-term outcomes, and be integrated into the health technology assessments, in particular relative effectiveness assessments (REAs) and cost-effectiveness assessments (CEAs) of novel or existing drugs in clinical practice, thereby supporting Randomized Clinical Trials (RCT) evidence.[19]
Real world data can be collected from different sources. According to a definition provided by the US Food and Drug Administration (FDA), RWD may include "data related to patient health status and/or the delivery of health care routinely collected from: electronic health records (EHRs), claims and billing data, [data from] product and disease registries, patient-generated data including home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices."[20]

5. Data Protection framework for processing RWD

As access to and use of RWD involves processing of sensitive health-related data collected from patients and research participants, it is crucial to investigate the lawful basis for processing RWD data and the relevant requirements when processing RWD for research purposes. Use of RWD may fall under the scope of relevant personal data protection regulations, if it is considered to be personal data. Notably, RWD may include both individual level patients' and research participants' data, but also aggregate data which would fall outside the scope of the GDPR. Processing individual level data that are considered as identifiable, consequently, needs to be on the basis of one of the lawful grounds recognized by the GDPR. In the Article 6 (1) of the GDPR the lawful grounds for processing personal data are listed, including obtaining consent, processing for compliance with a legal obligation, public interest, or legitimate interest. In addition, under Article 9 (2) of the GDPR, the special categories of data, including health data, can be processed on the basis of one of the recognized lawful bases, including obtaining explicit consent, among others. Furthermore, processing of special categories of data can be permitted if the processing is for scientific research purposes, and under the so-called research exemption provisions.
Processing RWD however may fall under various legal provisions, first depending on the source and a legal basis for the initial data collection and second, the specific purpose of using RWD. In the following, we will further elaborate these elements.

[19] A. MAKADY, et al. *Using Real-World Data in Health Technology Assessment (HTA) Practice: A Comparative Study of Five HTA Agencies,* in *Pharmacoeconomics*, 36, 3, 2018, 359-368, https://rdcu.be/ciOQb.
[20] FDA*, Real World Evidence*, 2020, available at: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

5.1. Source of the initial data collection and a question of further processing of data

As it has been briefly explained earlier, RWD may refer to the data that are extracted from diverse sources, including electronic health records (EHRs), claims and billing data, data from product and disease registries, patient-generated data including home-use settings, and data gathered from other sources that can inform on health status, such as mobile devices. In this sense, use of RWD could be mainly based on the processing of the existing data, rather than new data collection. Therefore, the data controllers should clarify whether processing of RWD deviates from the original data collection purposes, thus considered as further processing of data, and whether such further processing is allowed by law. The GDPR allows further processing under a number of conditions:

"The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allowed the collection of the personal data is required."[21]

One can argue that further processing of for example electronic health records or patient-generated data such as those gathered from mobile services, for clinical trials or HTA purposes may not immediately considered as compatible with purposes for which the personal data were initially collected. However, according to the GDPR, the further processing of data may also be regarded as compatible and lawful if the processing *"is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, Union or Member State law".*

A question here is whether processing RWD can be considered for the purpose of the "exercise of official authority vested in the controller, Union or Member State Law". Our answer here is mainly dependent on how the nature of RDW use has been defined. RWD can be used to inform the assessment of the reliability, safety or relative effectiveness of a treatment. Since in principle this is part of the objectives of clinical trials, the questions arise whether processing of RWD is to comply with the legal obligations to which the sponsor and/or the investigator is subject to under the relevant clinical trials regulations. Looking at the EU Clinical Trials Regulations 2014 and a relevant opinion issued by the European Data Protection Board, it appears that not all data processing in the framework of clinical trials are considered to be for the purposes of compliance with legal obligations. In fact, besides the processing that is strictly necessary for safety reporting or archiving obligation of clinical data, the rest of processing is more likely to be considered as a research activity. [22]

In case processing of RWD is considered to be for scientific research purposes, then the processing should be in compliance with the relevant requirements set by a number of regulations. While some of these provisions are embedded in the applicable data protection regulations, others are set by regulations concerning human subjects research and clinical research.

[21] Recital 50, GDPR.

[22] European Data Protection Board, Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR). Available online at: https://edpb.europa.eu/our-work-tools/our-documents/avis-art-70/opinion-32019-concerning-questions-and-answers-interplay_en

5.2. Legal requirements for using RWD for research purposes

To begin with, it is important to note that processing data for scientific research purposes is being recognized as a lawful basis for processing of special categories of data under Article 9(2), and a compatible purpose for further processing of data under Recital 50 of the GDPR. Notably, the recital 50 appears to assimilate purpose specification and lawfulness in the case of reuse for the purposes of scientific research. To further elucidate this recital, the European Data Protection Supervisor (EDPS) has stated that: "We [EDPS] would therefore argue that, in order to ensure respect for the rights of the data subject, the compatibility test under Article 6(4) should still be considered prior to the reuse of data for the purposes of scientific research, particularly where the data was originally collected for very different purposes or outside the area of scientific research. Indeed, according to one analysis from a medical research perspective, applying this test should be straightforward." [23]

Furthermore, processing data for scientific research purposes should be in accordance with Article 89 of the GDPR based on Union or Member States laws, and subject to adopting adequate technical and organizational measures to safeguard the rights of the data subjects. To date, diverse approaches are being adopted in implementation of Article 89 across the Member States, in particular in terms of safeguards required, when processing data for research purposes. Notably, further requirements may still apply due to the fact that collection and use of data for biomedical research and in particular clinical research concerns other applicable regulations regarding human subjects' research.

Finally, it might be difficult to draw a clear distinction between research and care when processing RWD data. In fact, the use of RWD may not be only necessary for generating evidence to benefit research purposes, but also improve health care decisions, or for the purposes of so-called learning healthcare systems. As Budrionis and Bellika put it, the learning healthcare system, among other things, focuses on "exploring the potential of data collected in daily clinical practice as a source of up-to-date minimally biased population-specific knowledge, which could be implemented into clinical practice in a more agile manner than randomized controlled trials."[24]

Taking such uses into considerations, it will be difficult to clearly separate research from clinical practice when using RWD. Notably, this poses extra complexities regarding feasibility of drawing a clear line between using data for research, care, quality assurance and proving safety and efficiency of medical products purposes when using RWD. In addition, the growing interest in linking various existing health-related databases, from electronic health records to data collected through wearables and apps, further challenges holding to traditional distinction between various purposes in order to define the legal requirements for processing data.

Furthermore, this has implications for determining the rights and responsibilities of the involved parties. This is particularly pertinent in some respects, such as the requirements for consent or return of secondary findings, where the applicable legal frameworks differ based on the nature of the respected activity.

[23] European Data Protection Supervisor- Preliminary Opinion on Data Protection and Scientific Research- 01/2020 & EDPB- Opinion No 3/2019 .

[24] A. BUDRIONIS, J. BELLIKA. *The Learning HealthCare System. Where are we now? A systematic Review,* in *Journal of Biomedical Informatics,* 64, 2016, 87-92.

ISSN 2284-4503

6. Path forward

Biomedical research is significantly benefiting from processing health and genomic data which are collected outside traditional research setting. Use of various mHealth application and online platforms, combined with access to other data sources such as EHR and patient registries has expanded biomedical researchers' access to a wide range of patient generated data and enables secondary uses of existing non-research databases. It is expected that this trend to be continued in the future and to be further facilitated by accessibility of various mHealth applications.

Processing of personal data for biomedical research purposes needs to be in compliance with the applicable data protection regulations. As we have shown above, processing personal data collected outside traditional research settings is associated with complexities regarding the nature of data processing and the relevant legal grounds and requirements for processing data. As we have seen in the case of data collection via mHealth applications and online platforms, the individuals may not be fully aware of secondary uses of data for research purposes. In view of the shortcomings of current privacy policies in adequately informing individuals about potential secondary uses of data, it is crucial to utilize innovative approaches to enhance transparency of the data processing for research purposes.

Further involving individuals in sharing genomic and health data, by enabling them to share their data directly with the interested parties seems advantageous, as this allows the individuals to have say in the way their data have been further processed for research purposes. However, we should note that "individual control" on health data or data ownership rhetoric in the context of health data are associated with significant limitations. In many jurisdictions, there is no legal framework for health data ownership for the individuals. Furthermore, individuals may not be adequately informed about the implications of exchanging their health data for monetary and non-monetary incentives such as tokens, shares or free sequencing. For instance, they may not be able to fully withdraw their consent for sharing data once they received free sequencing or the like in exchange for sharing data.

Furthermore, collection of data from various sources, may render it difficult to maintain to the traditional distinction of care-research when determining the applicable legal framework for data processing and clarifying the roles and responsibilities of the involved parties. As we have shown in this paper, for instance, the RWD can be collected both from existing clinical or patient-generated databases, to be used for various purposes including research purposes or informing the clinical decision-making. In that sense, such data processing may not strictly fit into one specific category. Therefore, rather than holding to traditional care-research distinction to determine the rights and responsibilities of the involved parties, it is crucial to identify the intricacies of data processing arising from secondary uses of data and adopt adequate safeguards in response to the associated risks. Tools and mechanisms such as Data Protection Impact Assessment (DPIA) which has been foreseen by the GDPR can be utilized in such assessment of the risks.

The next logical step would be adopting organizational and technical safeguards in the view of the identified risks. In this regard, taking advantage of emerging technological advancements in the area of data sharing and access such as distributed networks, which reduce a need for actual transfer of data is highly recommended. Furthermore, risks of re-identifiability of the individuals should be fully assessed and the adequate technical and organizational safeguards to be used to protect the data

subjects. Notably, the growing interest in connecting various research and clinical databases may pose new types of risks related to privacy and data protection.

Last but not least, when there is a question of further processing of data, which can lead to access to data by third parties, such as biotech or pharma companies, it is crucial to enhance transparency of data processing. To this end, the adequate information regarding the data processing should be communicated to the data subjects. Enhancing transparency when using patients and individual's data for research purposes would lead to higher trust on researchers and healthcare institutions.

ISSN 2284-4503